

DOCUMENT RESUME

ED 189 114

TM 800 277

AUTHOR Berk, Ronald A.
 TITLE Domain-Referenced Versus Mastery Conceptualization of Criterion-Referenced Measurement: A Clarification.
 PUB DATE Apr 80
 NOTE 19p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).
 EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS *Classification; *Criterion Referenced Tests; *Mastery Tests; Measurement Objectives; Models; *Test Construction; Test Format
 IDENTIFIERS *Domain Referenced Tests

ABSTRACT

Two approaches to criterion-referenced measurement are described and contrasted--domain-referenced testing and mastery testing. This paper is organized according to ten issues or stages in test construction: (1) content domain specification; (2) item construction; (3) item domain specification; (4) item analysis; (5) item selection; (6) parallel forms assumption; (7) cutting score selection; (8) test validity; (9) test reliability; and (10) score interpretation. (Author/GDC)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED189114

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

DOMAIN-REFERENCED VERSUS MASTERY CONCEPTUALIZATION
OF CRITERION-REFERENCED MEASUREMENT: A CLARIFICATION

Ronald A. Berk

The Johns Hopkins University

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

R. Berk

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American
Educational Research Association

Boston, April 1980

TM 800277

Abstract

Two major conceptualizations of criterion-referenced measurement seem to have emerged over the past decade--domain-referenced and mastery. This paper attempts to clarify these terms and synthesize the methodological characteristics emanating from both research and practice. The domain-referenced and mastery approaches are defined and analyzed according to 10 stages and/or issues in test development: (1) content domain specification, (2) item construction, (3) item domain specification, (4) item analysis, (5) item selection, (6) parallel forms assumption, (7) cutting score selection, (8) validity, (9) reliability, and (10) score interpretation.

Since the term "criterion-referenced measurement" was first applied to proficiency assessment by Glaser and Klaus (1962), more than 50 descriptions of a criterion-referenced test have appeared in the literature. Various comparisons of these descriptions suggest there is general agreement that the test is intended to reference an individual's score to a well-defined domain of behaviors (cf. Gray, 1978; Hambleton, Swaminathan, Algina, & Coulson, 1978; Livingston, 1977; Millman, 1974). Popham's (1978, p. 93) definition most accurately captures that conceptualization: A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain. Such tests are often referred to domain-referenced. The principal concern in their development is obtaining rigorous and precise domain specifications to maximize the interpretability of an individual's domain score.

An alternative conceptualization of criterion-referenced measurement derived from mastery learning theory (Block & Burns, 1976; Bloom, 1968; Carroll, 1963, 1970; Mayo, 1970) and popularized by many of the recent technical articles on item statistics (e.g., Berk, 1978b, 1980b) and reliability (e.g., Brennan & Kane, 1977; Subkoviak 1976) is represented by the mastery test. It is used to classify students as masters and nonmasters of an objective in order to expedite individualized instruction. Empirical item analysis procedures are recommended to determine whether the items are instructionally sensitive or discriminate between instructed and uninstructed groups. Methods for setting absolute performance standards for mastery, the estimation of classification errors, and the determination of mastery-nonmastery decision consistency are particularly important.

This paper is designed to clarify the terminology related to criterion-referenced measurement and synthesize the methodological characteristics emanating from both research and practice. The domain-referenced and mastery

conceptualizations are compared in terms of the test construction process and score interpretation.

Method

The domain-referenced and mastery conceptualizations of criterion-referenced measurement were analyzed according to 10 major stages and/or issues in test development: (1) content domain specification, (2) item construction, (3) item domain specification, (4) item analysis, (5) item selection, (6) parallel forms assumption, (7) cutting score selection, (8) validity, (9) reliability, and (10) score interpretation.

Comparison of Two Conceptualizations

The characteristics of the domain-referenced and mastery conceptualizations are summarized in Table 1. A discussion of these characteristics follows.

Insert Table 1 about here

Content Domain Specification

Domain-referenced. The primary emphasis in the development of a domain-referenced test is at this first stage. The "unambiguous" definition of a behavioral domain and the generation of an item domain from that definition have been recommended to maximize the interpretability of an individual's domain score. These goals are difficult to achieve even with the technologies currently available (cf. Popham, 1980). Among the various domain specification strategies, only item transformations, item forms, and algorithms come close to attaining "unambiguous" definitions via sophisticated rule structures (Berk, in press).

Mastery. The mastery test conceptualization is built on a set of instructional and behavioral objectives. While the objectives-based definition is relatively less precise and "ambiguous" compared to the preceding strategies, it is extremely practicable. In addition, it is possible to improve upon this definition by using amplified objectives, IOX test specifications, or mapping sentences.

Item Construction

Domain-referenced. The domain-referenced approach requires detailed, explicit rules for generating the test items. The "unambiguous" specification strategies noted above provide objective, computer-based technologies for producing item domains and sampling items from those domains (Millman, 1980).

Mastery. The objectives-based strategies associated with the mastery conceptualization necessitate a heavy reliance upon traditional item construction rules. The use of mapping sentences (facet theory) in conjunction with objectives (Berk, 1978a), however, holds considerable promise for refining the item generation process so that it is more systematic and mechanical than simply applying the traditional rules.

Item Domain Specification

Domain-referenced. Until recently the distinction between infinite and finite item domains was rarely considered in the test development process. The technological advances over the past decade in the specification of behavioral domains, however, have made it possible to actually generate "all possible items" in a domain. The level of precision in the specifications necessary to produce a finite domain is quite unlike that of the objectives-based schemes. There are at least two advantages to the exact specification and enumeration of the items in a domain: (1) the items that comprise one test form or alternate forms can be drawn from the domain using a random or stratified

random sampling plan and (2) the characteristics of the domain can be identified and, consequently, can be incorporated into the sampling design, e.g., stratification by difficulty level, and other stages of test construction and score analysis, e.g., parallel forms assumption, selection of an appropriate reliability index.

Mastery. The objectives employed to define the content domain for a mastery test may suggest either type of item domain. However, typically such a distinction is theoretical except in the case of very specific behavioral objectives; for example, the student will be able to multiply all combinations of single digit numbers. The items written to be included on a particular test are usually not sampled from any domain. Rather, those items viewed collectively are assumed to constitute a sample from the domain of "all possible items" that could have been written from the objectives. Since the distinction between item domains makes little practical difference in the context of mastery test construction, the concept of "all possible items" is often associated with the assumption of an infinite item domain.

Item Analysis

Domain-referenced. An item analysis is unnecessary and, in fact, undesirable for domain-referenced test items. The use of item statistics for refining and selecting the items would theoretically destroy the defining character of the test, thereby weakening the interpretability of the domain score (Millman & Popham, 1974). Furthermore, given the precision of most of the strategies that are used to generate the item domain, the need to assess item-objective congruence or search for flaws in the items seems questionable.

Mastery. An item analysis is desirable for mastery test items due to the intended uses of the test scores for mastery-nonmastery decisions and to the traditional, relatively subjective procedures employed in writing the items

(Berk, 1980b). The items must be scrutinized to appraise whether they measure their respective objectives and differentiate between groups of masters and nonmasters of those objectives (item discrimination or instructional sensitivity). In addition, they must be free of structural flaws that could cue or confuse the students.

Item Selection

Domain-referenced. As indicated previously the items in a domain-referenced test are sampled randomly from the item domain. This assures that they are representative and congruent with the domain.

Mastery. In contrast, mastery test items are selected nonrandomly based on their congruence with the objectives and their sensitivity to instructional treatments. The item selection criteria recently presented by Berk (1980b) emphasize clearly the primary importance of congruence and the secondary importance of statistical item properties.

Parallel Forms Assumption

The assumption about parallel forms of a criterion-referenced test is related to (1) how the item sample that comprises one test form and the samples that comprise alternate forms are generated and (2) the characteristics of the item domain from which those samples are generated.

Domain-referenced. Since domain-referenced item samples can be drawn randomly from an item domain, they are said to be randomly parallel. No specific statistical properties apply to these test forms nor is there any condition imposed that limits the variability of item difficulty levels in the domain. When the test maker cannot actually build test forms by randomly sampling from the domain, those forms and their characteristics can be "assumed" (cf. Cardinet, Tourneur, & Allal, 1976). As Brennan (1980) indicates:

The assumption of randomly parallel tests is a useful idealization. Although the items of a particular test may not have been drawn at random, one can usually conceive of a universe of items from which the test items might have been drawn. (p. 226)

Mastery. When items are written from objectives-based specifications, it is possible to develop two or more item samples for the same objective and domain. Frequently, those samples are constructed systematically to contain content and to yield means, variances, and item intercorrelations identical to the first sample. Item samples or alternate test forms with those properties are said to be classically parallel or equivalent to the first form. This statistical equivalence imposes restrictions on the domain. It requires that the content or set of behaviors being measured by the items be relatively homogeneous and, therefore, that the items in the domain possess similar difficulty levels. Very often when a test maker is not able to produce parallel forms, those forms and their statistical properties are simply "assumed."

Cutting Score Selection

Domain-referenced. Although a domain-referenced test is not designed expressly for mastery-nonmastery decision making, a cutting score may be set. The approaches currently available for determining a mastery standard have been organized by Hambleton and Eignor (1979) into three categories: (1) judgmental, (2) empirical, and (3) judgmental-empirical combination. Detailed descriptions of the approaches have been compiled by Hambleton (1980) and Hambleton, Powell, and Eignor (1979). The choice of a single approach appears to be dependent on the specific test application and several factors related to that application, including political considerations, decision consequences, practical constraints, statistical expertise

and resources, and the types of mastery decision to be made.

Mastery. In the case of mastery testing a standard must be set. The number of cutting scores that need to be selected for a test will differ with the structure of the test and the types of decision. For objectives-based tests used for individual decisions at the classroom level, cutting scores should be chosen for each item cluster or subtest keyed to an instructional objective. There is nothing implicit in any of the approaches that dictates that the cutting score has to be identical for all subtests. The customary use of one "blanket mastery standard" for all objectives seems inadvisable for the following reasons. First, objectives vary in level of complexity and, therefore, applying the same standard to different objectives will be insensitive to that variability. Second, the difficulty and discrimination indices of items measuring one objective may be quite different than those for other objectives; consequently, it is unrealistic to expect a high proportion of students to obtain the same standard for each objective. Third, since the instruction associated with different objectives is rarely given the same time allotment and attention, performance on some objectives might be noticeably better than the performance on others. Again, this is contradictory with setting the same standard for all objectives.

If the mastery test happens to be developed as a "minimum competency test" that is used for individual certification decisions at the school level, a single cutting score may be set for the total test rather than for each objective. Decisions regarding graduation and grade-to-grade promotion ordinarily do not require objective-by-objective analysis, although it is possible to specify a certain number of objectives that must be mastered in order to "pass." The information on objective mastery might be employed for follow-up diagnostic purposes along with instructional prescriptions for students who "fail."

Validity

While the assessment of content, criterion-related, and construct validity is applicable to both test conceptualizations (cf. Hambleton, 1980; Linn, 1979b),

the emphases differ markedly. These emphases are discussed next.

Domain-referenced. The tasks involved in building a domain-referenced test concentrate, for the most part, on establishing content validity. The rigor of the domain specifications and its direct link with the test items are features uncommon to existing achievement tests. The effort devoted in the initial stages is designed to pay off with the dividends of descriptive clarity and precise score interpretations (cf. Popham, 1980, pp. 25-26). Procedures that can be used to improve the content validity of the test have been devised by Hambleton (1980).

Another important concern is the construct validity of the scores. Empirically-based construct validity studies that employ factor analysis or Guttman scalogram analysis should be conducted. Hambleton (1980) has identified one particular study of functional literacy test scores by Kirsh and Guthrie (1979) that he considers exemplary.

Mastery. Given the objectives-based specifications of the mastery test, a special review process becomes necessary to determine content validity. Hambleton (1980) urges an examination of the validity, technical quality, and representativeness of the items. Techniques for evaluating item-objective congruence using rating scales and matching tasks have been suggested. While this review usually will occur at the item analysis stage, it is significant to note that the judgmental and empirical procedures used to establish content validity for the mastery test are quite different from the domain specification strategies used to build in content validity for the domain-referenced test.

In addition to the steps required to insure content validity, it is also essential to investigate the criterion-related and construct validity of the test. The mastery decision framework provides a specific focus for such investigations. The most complex and problematic aspect of this line of inquiry pertains to the definition of the success or mastery criterion. An operational definition must be specified and criterion success and nonsuccess groups of persons need to be

identified. The criterion-related validity studies may be concurrent or predictive depending upon how success is defined, e.g., success in the current program, success in the next unit of instruction.

Furthermore, a particular type of construct validity for mastery test scores that deserves special attention is decision validity. This has also been called discriminant validity. It relates to the evidence gathered on the accuracy of mastery-nonmastery classification decisions (Hambleton, 1980, pp. 97-98). That evidence is expressed often as probabilities of correct and incorrect (false mastery and false nonmastery) classifications. Such evidence is crucial in the attempt to justify the selection of the cutting score. When decision validity evidence is not available, it seems pointless even to compute a reliability index. A high reliability index based on an "invalid" or "unjustified" standard, for example, might mean that the mastery test can consistently classify students into the wrong groups. Consistent decision making without accurate decision making has questionable value in criterion-referenced evaluation.

Reliability

The types of reliability that are appropriate for domain-referenced and mastery tests are contingent upon the parallel forms assumption, cutting score selection, and the intended uses of the test scores. The reviews of reliability indices by Berk (1980a), Hambleton, Swaminathan, Algina, and Coulson (1978), Linn (1979a), and Shepard (1980) suggest specific guidelines for selection and use.

Domain-referenced. Since the major purpose of the domain-referenced test is to estimate and to clearly interpret an individual's domain score in terms of a well-defined behavioral domain, point and interval estimation approaches would seem to be most apropos. When that purpose is viewed in conjunction with the randomly parallel test forms assumption, there appear

to be at least two point estimates and two standard errors of measurement for interval estimates that a test maker should consider (see Berk, 1980). In addition, Brennan (1980) has proposed two indices of "dependability" derived from generalizability theory that also merit serious attention. One of those indices $\phi(\lambda)$ is a function of the cutting score. When it is desirable to make mastery-nonmastery decisions with a domain-referenced test, a cutting score should be selected and the $\phi(\lambda)$ index could be used as a measure of test score dependability relative to λ .

Mastery. Given the usual though not essential assumption of classically parallel test forms, the selection of a cutting score, and the use of the scores for mastery-nonmastery decisions, different approaches to mastery test reliability are indicated. The classification consistency estimates of p_o or κ based on a threshold loss function have been highly recommended. The recent comparisons of various one- and two-administration estimates suggest that Hambleton and Novick's (1973) method for estimating p_o and Huynh's (1976) method for estimating p_o and κ may be especially useful (Berk, 1980a; Subkoviak, 1980). When distinctions among degrees of mastery and nonmastery along the score continuum are desirable, not just the qualitative master-nonmaster classification assumed by the threshold loss function, Livingston's $k^2(X, T_x)$ squared-error loss index should be considered.

Score Interpretation

Domain-referenced. The estimate of a student's domain score supplies information about how much the student knows or his/her level of skill in a given content domain. It is difficult to make placement, formative, or summative decisions, however, based on that information alone. Some reference point or standard is often required. Once the score is referenced to a mastery standard, individual decision making becomes considerably easier.

For program evaluation decisions that focus usually on the effective-

ness of curricular materials, instructional methods, instructional media, or any combination of materials, methods, and media, the use of the average domain score is appropriate. Estimates of average domain scores vary as a function of different matrix sampling designs (see Sirotnik, 1974).

Mastery. The decisions made by the classroom teacher within the context of mastery learning theory and diagnostic-prescriptive teaching have best demonstrated the utility of mastery test scores. Typically, those decisions are derived from an objectives-based assessment of individual strengths (mastery) and weaknesses (nonmastery). They involve placing students at the appropriate points in the instructional sequence, monitoring students' successes and failures as they progress through the program, and assigning letter grades/pass-fail to certify competencies in a subject area or course. Recently the latter type of summative evaluation has been extended in the form of minimum competency certification for grade-to-grade promotion and high school graduation.

In addition to these different types of individual decisions, the scores may be used to assess program effectiveness at the classroom, school, district, and state levels. The actual decisions may be to continue, to modify, or to terminate a given program. In order to supply information that will be meaningful for these decisions it is often necessary to summarize the test score results into various forms such as the percentage of students demonstrating mastery of each objective and the percentage of students demonstrating mastery of sets of objectives. The specific summary will differ according to the level of the decision maker and the nature of the decision. The summary that is needed for a particular decision is composed by aggregating the individual test scores at the appropriate decision level.

A Final Comment

This paper was intended to clarify two major conceptualizations of a criterion-referenced test that seemed to have emerged from the research and practices over the past decade. Perhaps it raised more questions than it answered. However, it is hoped that the structure and accompanying discussion will help eliminate some of the apparent confusion over measurement terminology and will facilitate the appropriate use of the methodologies consistent with the particular conceptualizations. The delineation of alternative strategies or procedures at each stage of the test development process should also provide a guide for the increased and improved utilization of the available criterion-referenced measurement technology by teachers, evaluators, and test publishers.

References

- Berk, R. A. The application of structural facet theory to achievement test construction. Educational Research Quarterly, 1978, 3, 62-72. (a)
- Berk, R. A. A consumers' guide to criterion-referenced test item statistics. Measurement in Education, 1978, 9, 1-8. (b)
- Berk, R. A. A consumers' guide to criterion-referenced test "reliability." Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980. (a)
- Berk, R. A. Item analysis. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 49-79. (b)
- Berk, R. A. A comparison of six content domain specification strategies for criterion-referenced tests. Educational Technology, in press.
- Block, J. H., & Burns, R. B. Mastery learning. In L. S. Shulman (Ed.), Review of Research in Education (Vol. 4). Itasca, IL: F. E. Peacock, 1976.
- Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 1, No. 1.
- Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 186-232.
- Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.
- Cardinet, J., Tourneur, Y., & Allal, L. The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135.
- Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.
- Carroll, J. B. Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 48, 71-80.
- Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), Psychological principles in systems development. New York: Holt, Rinehart and Winston, 1962. Pp. 419-474.
- Gray, W. M. A comparison of Piagetian theory and criterion-referenced measurement. Review of Educational Research, 1978, 48, 223-249.
- Hambleton, R. K. Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 80-123.
- Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard-setting. In R. M. Jaeger and C. K. Tittle (Eds.), Minimum competency achievement testing. Berkeley, CA: McCutchan, 1979.

- Hambleton, R. K., & Novick, M. R. Toward an intergration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.
- Hambleton, R. K., Powell, S., & Eignor, D. R. Issues and methods for standard-setting. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.
- Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.
- Kirsh, I., & Guthrie, J. T. Toward a construct validity of functional reading tasks. Newark, DE: International Reading Association, 1979.
- Linn, R. L. Issues of reliability in measurement for competency-based programs. In M. A. Bunda and J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979. Pp. 90-107. (a)
- Linn, R. L. Issues of validity in measurement for competency-based programs. In M. A. Bunda and J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979. Pp. 108-123. (b)
- Livingston, S. A. Psychometric techniques for criterion-referenced testing and behavioral assessment. In J. D. Cone and R. P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1977. Pp. 308-329.
- Mayo, S. T. Mastery learning and mastery testing. Measurement in Education, 1970, 1, 1-4.
- Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974. Pp. 311-397.
- Millman, J. Computer-based item generation. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 32-43.
- Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.
- Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.
- Popham, W. J. Domain specification strategies. In R. A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 15-31.

Shepard, L. A. Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of research in education (Vol. 8). Itasca, IL: F. E. Peacock, 1980.

Sirotnik, K. A. Introduction to matrix sampling for the practitioner. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974. Pp. 451-529.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M. J. Decision-consistency approaches. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 129-185.

Table 1

**Characteristics of Domain-Referenced and Mastery Conceptualizations
of a Criterion-Referenced Test**

Stages/Issues in Test Development	Alternative Conceptualizations	
	Domain-Referenced	Mastery
1. Content domain specification	Possible strategies: a. Amplified objectives b. IOX test specifications c. Mapping sentences d. Item transformations e. Item forms f. Algorithms	Instructional and behavioral objectives
2. Item construction	Generation rules	Traditional rules
3. Item domain specification	Infinite or finite (contingent upon precision of domain specification strategy)	Infinite (or finite)
4. Item analysis	- - - - -	Item-objective congruence Item statistics (DIFF and DIS)
5. Item selection	Random	Nonrandom
6. Parallel forms assumption	Randomly parallel	Classically parallel
7. Cutting score selection	Optional	Required
8. Validity	Content Construct (decision)	Content Criterion-related Construct (decision)
9. Reliability	Possible approaches: a. Squared-error loss estimate of $\theta(\lambda)$ or θ b. Point estimate of domain score c. Interval estimate of domain score (individual specific or group specific standard error of measurement)	Possible approaches: a. Threshold loss estimate of p_o or κ b. Squared-error loss estimate of $k^2(X, T_x)$
10. Score interpretation ^a	Reference score to domain Reference score to cutting score for mastery-nonmastery decisions	Reference score to cutting score for mastery-nonmastery decisions

^a While score interpretation is not a stage or issue in test development per se, it is a major consideration in all of the preceding nine stages.